JOURNAL OF NONPARAMETRIC STATISTICS
https://doi.org/10.1080/10485252.2024.2313137

Taylor & Francis
Taylor & Francis Group

Robust sufficient dimension reduction via α-distance covariance

Hsin-Hsiung Huang[a], Feng Yu[b] and Teng Zhang[c]

[a]Department of Statistics and Data Science, University of Central Florida, Orlando, FL, USA; [b]Department of Mathematics, University of Minnesota Twin Cities, Minneapolis, MN, USA; [c]Department of Mathematics, University of Central Florida, Orlando, FL, USA

ABSTRACT

We introduce a novel sufficient dimension-reduction (SDR) method which is robust against outliers using α-distance covariance (dCov) in dimension-reduction problems. Under very mild conditions on the predictors, the central subspace is effectively estimated and model-free without estimating link function based on the projection on the Stiefel manifold. We establish the convergence property of the proposed estimation under some regularity conditions. We compare the performance of our method with existing SDR methods by simulation and real data analysis and show that our algorithm improves the computational efficiency and effectiveness.

ARTICLE HISTORY
Received 13 August 2023
Accepted 23 January 2024

KEYWORDS
α-distance covariance; central subspace; sufficient dimension reduction (SDR); manifold learning; robust statistics

1. Introduction

In regression analysis, sufficient dimension reduction (SDR) provides a useful statistical framework to analyse a high-dimensional dataset without losing any information. It finds the fewest linear combinations of predictors that capture a full regression relationship. Let Y be an univariate response and $X = (x_1, \ldots, x_p)^T$ be a $p \times 1$ predictor vector, SDR aims to find a $p \times d$ matrix $\boldsymbol{\beta}$ such that

$$Y \perp\!\!\!\perp X \mid \boldsymbol{\beta}^T X$$

which denotes the statistical independence.

Sufficient dimension reduction (SDR) based on the conditional distribution of the response (Li 1991; Cook and Weisberg 1991; Xia et al. 2002; Yin and Li 2011) provides the reduced predictors without loss of regression information. Recently, SDR methods using distance covariance (dCov) have been developed (Sheng and Yin 2013, 2016), and such methods do not need a constant covariance condition and distribution assumptions on X, $X \mid Y$ or $Y \mid X$. Therefore, it has broad applications for continuous and discrete variables from various distributions. Several robust sufficient dimension reduction methods have proposed for coefficient estimation such as the robust sufficient dimension reduction using the ball covariance (Zhang and Chen 2019) and the expected likelihood based method that

CONTACT Hsin-Hsiung Huang hsin.huang@ucf.edu Department of Statistics and Data Science, University of Central Florida, Orlando, Florida 32816, USA

minimises the Kullback–Leibler distance (Yin and Cook 2005; Zhang and Yin 2015). In this article, we propose a robust method of sufficient dimension reduction via the α-distance covariance (α-dCov) between the response and the predictors and develop a new algorithm for estimating directions in general multiple-index models with a form

$$Y = g(\boldsymbol{\beta}^T X, \epsilon),$$

where g is an unknown link function (Yin et al. 2008; Xia 2008; Sheng and Yin 2013). The rest of this article is organised as follows: Section 2 describes our robust α-dCov method and a corresponding outlier detection method, including motivation, theoretical results, the estimation algorithm and testing procedure. We introduce the consistency theorem in Section 3. Section 4 contains simulation and real data studies. We summarise our work in Section 5.

1.1. Generalised distance covariance

Distance covariance (Székely et al. 2007) is a popular dependence measure for two random vectors of possibly different dimensions and types. In recent years, there have been concentrated efforts in the literature to understand the distributional properties of the sample distance covariance in a high-dimensional setting, with an exclusive emphasis on the null case that X and Y are independent. Distance covariance can be generalised to include powers of the Euclidean distance. Define

$$\begin{aligned}\nu^2(X, Y; \alpha) := {} & \mathrm{E}[\|X - X'\|^\alpha \, \|Y - Y'\|^\alpha] + \mathrm{E}[\|X - X'\|^\alpha] \, \mathrm{E}[\|Y - Y'\|^\alpha] \\ & - 2\mathrm{E}[\|X - X'\|^\alpha \, \|Y - Y''\|^\alpha], \end{aligned} \tag{1}$$

where (X, Y), (X', Y'), (X'', Y'') are independent and identically distributed (i.i.d.) with respect to the joint distribution of (X, Y) (Székely and Rizzo 2014). As discussed in Székely et al. (2007), for every $0 < \alpha < 2$, X and Y are independent if and only if $\nu^2(X, Y; \alpha) = 0$. When $\alpha = 1$, it reduces to the classical distance covariance. When $0 < \alpha < 1$, it can be considered as a more robust version of distance covariance as it reduces the influence of large values of $\|X - X'\|$, $\|Y - Y'\|$, and $\|Y - Y''\|$ that might be contributed to outliers.

1.2. Central space estimation via α-dCov

Let $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i) : i = 1, \ldots, n\}$ be n random samples from random variables (X, Y). In addition, $\mathbf{X}$ denotes a $p \times n$ data matrix whose columns are $X_1, \ldots, X_n$ and $\mathbf{Y} = [Y_1, \ldots, Y_n]$ denotes a $1 \times n$ response data matrix. In this article, we consider univariate responses. However, the method can naturally be extended to multivariate responses without any issue due to the nature of α-dCov. The empirical solution of the SDR method based on α-dCov for these n observations relies on solving the following objective function (Székely et al. 2007; Sheng and Yin 2016):

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^{p \times d}} \nu_n^2(\boldsymbol{\beta}^T \mathbf{X}, \mathbf{Y}, \alpha). \tag{2}$$

with constraint $\boldsymbol{\beta}^T \Sigma_X \boldsymbol{\beta} = \mathbf{I}_d$ and $1 \leq d \leq p$, where ν_n is the empirical version of ν defined in Equation (1).

The empirical distance dependence statistics ν_n is defined as follows. For $k, l = 1, \ldots, n$, we compute the Euclidean distance matrices $(a_{kl}) = (|X_k - X_l|_p^\alpha)$ and $(b_{kl}) = (|Y_k - Y_l|^\alpha)$ for $0 < \alpha < 2$ (Székely and Rizzo 2009). Define

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}, \quad k, l = 1, \ldots, n,$$

where

$$\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \quad \bar{a}_{\cdot l}, = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}.$$

Similarly, we define $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$, for $k, l = 1, \ldots, n$. The nonnegative sample distance covariance $\nu_n(\mathbf{X}, \mathbf{Y})$ and sample distance correlation $R_n(\mathbf{X}, \mathbf{Y})$ are defined by

$$\nu_n^2(\boldsymbol{\beta}^T\mathbf{X}, \mathbf{Y}, \alpha) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl} \tag{3}$$

and

$$R_n^2(\mathbf{X}, \mathbf{Y}, \alpha) = \begin{cases} \frac{\nu_n^2(\mathbf{X},\mathbf{Y},\alpha)}{\nu_n^2(\mathbf{X},\alpha)\nu_n^2(\mathbf{Y},\alpha)}, & \text{if } \nu_n^2(\mathbf{X},\alpha)\nu_n^2(\mathbf{Y},\alpha) > 0; \\ 0, & \text{if } \nu_n^2(\mathbf{X},\alpha)\nu_n^2(\mathbf{Y},\alpha) = 0, \end{cases}$$

respectively, where the sample distance variance is defined by

$$\nu_n^2(\mathbf{X}, \alpha) := \nu_n^2(\mathbf{X}, \mathbf{X}, \alpha) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2.$$

Following Wu and Chen (2021), we have the following equivalence. Let $\mathbf{C} = \hat{\Sigma}_X^{\frac{1}{2}}\boldsymbol{\beta}$ and $\mathbf{Z} = \hat{\Sigma}_X^{-\frac{1}{2}}\mathbf{X}$, the target function (2) can be rewritten as

$$\max_{\mathbf{C}} \nu_n^2(\mathbf{C}^T\mathbf{Z}, \mathbf{Y}, \alpha) := \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}(\mathbf{C})B_{kl}, \text{ s.t. } \mathbf{C} \in \mathrm{St}(d, p), \tag{4}$$

where $a_{kl}(\mathbf{C}) = \|\mathbf{C}^T Z_k - \mathbf{C}^T Z_l\|^\alpha$. We use the same notation $\mathrm{St}(d, p) = \{\mathbf{C} \in \mathbb{R}^{p\times d} \mid \mathbf{C}^T\mathbf{C} = I_d\}$ with $d \le p$ is referred to the Stiefel manifold and $T_{\mathbf{C}}\mathrm{St}(d, p)$ is the tangent space to $\mathrm{St}(d, p)$ at a point $\mathbf{C} \in \mathrm{St}(d, p)$. We assume that $Y = g(\mathbf{C}^T\mathbf{Z}, \epsilon)$, where $\mathbf{C}$ is a $p \times d$ matrix, ϵ is an unknown random error independent of $\mathbf{Z}$, and g is an unknown link function. We propose a new method to estimate a basis of the central subspace $S_{Y|Z} = \mathrm{Span}(\mathbf{C})$ and denote $\nu_n^2(\mathbf{C}^T\mathbf{Z}, \mathbf{Y}, \alpha)$ as $F(\mathbf{C})$.

2. Algorithm

We develop an iterative algorithm based on the gradient descent algorithm on the Stiefel manifold. Here P_S is a projection on the Stiefel manifold (Dalmau-Cedeno and Oviedo 2017). By Proposition 3.4 (the projection onto Stiefel manifolds) of Absil and Malick (2012), we let $\bar{\mathbf{C}} \in \mathrm{St}(d, p)$ for any $\mathbf{C}$ such that $\|X - \bar{X}\| < \sigma_1(\bar{\mathbf{C}})$, where $\sigma_1(\bar{\mathbf{C}})$ is the largest singular value, then the projection of $\mathbf{C}$ onto $\mathrm{St}(d, p)$ exists uniquely, and

Algorithm 1 rSDR: robust SDR

1: **Input:** The samples $\{(y_i, \mathbf{Z}_i), i = 1, \ldots, n\}$, initial $\mathbf{C}^{(0)}$.
2: **Initialisation:** $\mathbf{C}^{(0)}$.
3: **for** iter $= 0, 1, \ldots$ **do**
4: Let $\mathbf{C}^{(\text{iter}+1)} = P_S(\mathbf{C}^{(\text{iter})} + \alpha_1^{(\text{iter})} \partial_{\mathbf{C}} F(\mathbf{C}^{(\text{iter})}))$ or $\mathbf{C}^{(\text{iter}+1)} = P_S(\mathbf{C}^{(\text{iter})} + \alpha_1^{(\text{iter})} \partial_{\mathbf{C}} F(\mathbf{C}^{(\text{iter})})(\mathbf{I} - \mathbf{C}^{(\text{iter})\,T}\mathbf{C}^{(\text{iter})}))$, where $P_S(\cdot)$ is the projection on the Stiefel manifold and $\frac{\partial}{\partial \mathbf{C}} F(\mathbf{C})$, and $\alpha_1^{(\text{iter})}$ is chosen by a line search.
5: Repeat steps 4 until $\|F(\mathbf{C}^{(\text{iter})}) - F(\mathbf{C}^{(\text{iter}-1)})\|_F \leq \epsilon_n$ where ϵ_n is a pre-specified threshold, or the number of iterations exceeds the upper limit: iter $> N^{(\max)}$.
6: **end for**
7: **Output:** Estimated coefficients $\hat{\mathbf{C}}$.

can be expressed as $P_S(\mathbf{C}) = \sum_{i=1}^{d} u_i v_i^T$, given by a singular value decomposition of $\mathbf{C}$. Alternatively, let the SVD of $\mathbf{C} \in \mathbb{R}^{p \times d}$ be $\mathbf{C} = \mathbf{U}\Sigma\mathbf{V}$, then $P_S(\mathbf{C}) = \mathbf{U}\mathbf{V}^T$.

Now we derive the explicit formula for $\partial_{\mathbf{C}} F(\mathbf{C})$, where $F(\mathbf{C}) = \nu_n^2(\mathbf{C}^T\mathbf{Z}, Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}(\mathbf{C}) B_{kl}$. Recall that $a_{kl}(\mathbf{C}) = \|\mathbf{C}^T\mathbf{Z}_k - \mathbf{C}^T\mathbf{Z}_l\|^\alpha$, the gradient is

$$\partial_{\mathbf{C}} F(\mathbf{C}) = \frac{\alpha}{n^2} \sum_{k,l=1}^{n} \frac{\mathbf{C}^T(\mathbf{Z}_k - \mathbf{Z}_l)(\mathbf{Z}_k - \mathbf{Z}_l)^T}{\|\mathbf{C}^T\mathbf{Z}_k - \mathbf{C}^T\mathbf{Z}_l\|^{2-\alpha}} B_{kl}, \tag{5}$$

and one may perform the manifold gradient descent algorithm as follows:

$$\mathbf{C}^{(\text{iter}+1)} = P_S\Big(\mathbf{C}^{(\text{iter})} + \alpha_1^{(\text{iter})} \partial_{\mathbf{C}} F(\mathbf{C}^{(\text{iter})})(\mathbf{I} - \mathbf{C}^{(\text{iter})}\mathbf{C}^{(\text{iter})\,T})\Big).$$

We remark that while there are various advanced Stiefel manifold optimisation algorithms such as the ones based on the Cayley transform (Wen and Yin 2013; Zhu et al. 2019) or geodesics (Absil et al. 2009), we applied the standard projected gradient descent algorithm as it is simpler to implementation and has the same order of computational cost per iteration of $O(p^2 d)$.

Implementation issues When implementing our approach, practical challenges may arise due to the potential for an extremely small denominator in Equation (5), disproportionately amplifying the influence of the (k, l)th term. To preemptively address this concern, we introduce a small positive regularisation parameter, denoted as η. Subsequently, we employ a regularisation technique on the objective function $F(\mathbf{C})$ such that the (k, l)th term of the gradient in Equation (5) remains bounded. In particular, we apply it to the regularised objective function denoted as $F_\eta(\mathbf{C})$:

$$F_\eta(\mathbf{C}) = \frac{1}{n^2}(\|\mathbf{C}^T Z_k - \mathbf{C}^T Z_l\|^2 + \eta)^{\alpha/2} B_{kl},$$

which leads to the regularised gradient formulation expressed as follows:

$$\frac{\alpha}{n^2} \partial_{\mathbf{C}} F_\eta(\mathbf{C}) = \sum_{k,l=1}^{n} \frac{\mathbf{C}^T(\mathbf{Z}_k - \mathbf{Z}_l)(\mathbf{Z}_k - \mathbf{Z}_l)^T}{(\|\mathbf{C}^T\mathbf{Z}_k - \mathbf{C}^T\mathbf{Z}_l\|^2 + \eta)^{(2-\alpha)/2}} B_{kl}. \tag{6}$$

3. Consistency theory

We consider a model with a general noise term

$$\mathbf{Y} = g(\boldsymbol{\beta}_0^T \mathbf{X}, \epsilon) = g(\mathbf{C}_0^T \mathbf{Z}, \epsilon),$$

where $\boldsymbol{\beta}_0$ is a $p \times d$ orthogonal matrix, $g(\cdot)$ is an unknown link function, $\mathbf{C}_0 = \hat{\boldsymbol{\Sigma}}_X^{\frac{1}{2}} \boldsymbol{\beta}_0$, and $\mathbf{Z} = \hat{\boldsymbol{\Sigma}}_X^{-\frac{1}{2}} \mathbf{X}$, and ϵ is independent of $\mathbf{Z}$. This model includes the model from Xia et al. (2002) that

$$\mathbf{Y} = g(\boldsymbol{\beta}_0^T \mathbf{X}) + \epsilon$$

as a special example.

Following Sheng and Yin (2016), we have the asymptotic properties of the estimator $\hat{\mathbf{C}}$ that is consistent. The statement and the proof is similar to that of Sheng and Yin (2013). It requires an additional assumption that depends on the decomposition of X into two independent components, and some discussions on this condition are available in Sheng and Yin (2013, Section 3.2). For example, it is satisfied when X is normal (Zhang and Yin 2015). In addition, this assumption also holds asymptotically when p is large (Hall and Li 1993).

The following proposition establishes the asymptotic properties of our estimator $\mathbf{C}$ up to some rotation matrix $\mathbf{Q}$. This implies the asymptotic property of the estimated central subspace as it is invariant to the rotation matrix.

Proposition 3.1: *Let $\mathbf{C} \in \mathbb{R}^{d \times p}$ be a basis of the central subspace $S_{Y|X}$ with $\mathbf{C}^T \boldsymbol{\Sigma}_X \mathbf{C} = I_d$. Suppose $P_{\mathbf{C}(\boldsymbol{\Sigma}_X)}^T X \perp\!\!\!\perp Q_{\mathbf{C}(\boldsymbol{\Sigma}_X)}^T X$ and the support of $X \in \mathbb{R}^{d \times p}$, say S, is a compact set. In addition, assume that there exists $\mathbf{C}' \in \mathbb{R}^{(p-d) \times p}$ such that $[\mathbf{C}, \mathbf{C}']^T \boldsymbol{\Sigma}_X [\mathbf{C}, \mathbf{C}'] = I_p$ and $\mathbf{C}^T X$ is independent of $\mathbf{C}'^T X$. Let $\hat{\mathbf{C}} = \arg\min_{\mathbf{C}^T \boldsymbol{\Sigma}_X \mathbf{C} = I_d} \nu_n^2(\mathbf{C}^T \mathbf{X}, \mathbf{Y})$, then there exists a rotation matrix $\mathbf{Q}$: $\mathbf{Q}^T \mathbf{Q} = I_d$ such that $\hat{\mathbf{C}} \xrightarrow{P} \mathbf{C}\mathbf{Q}$ (convergence in probability) as $n \to \infty$.*

Proof: Following Székely and Rizzo (2009, (4.1)), we have that for random variables X and Y from $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$,

$$\nu^2(X, Y, \alpha) = C \int_{t,s} \frac{|f_{X,Y}(t,s) - f_X(f) f_Y(t)|^2}{\|t\|^{p_1+\alpha} \|s\|^{p_2+\alpha}} \mathrm{d}t\mathrm{d}s,$$

where $f_X, f_Y, f_{X,Y}$ represent the characteristic functions of X, Y, and (X, Y) respectively.

The rest follows from the proof of Proposition 1 in Zhang and Yin (2015). For any $\boldsymbol{\beta} \neq \mathbf{C}$ that satisfies $\boldsymbol{\beta}^T \boldsymbol{\Sigma}_X \boldsymbol{\beta} = \mathbf{I}$, let $\boldsymbol{\beta}_1$ be the projection of $\boldsymbol{\beta}$ to the subspace spanned by $\mathbf{C}$ with an inner product induced by $\boldsymbol{\Sigma}_X$ (that is, $\boldsymbol{\Sigma}_X^{0.5} \boldsymbol{\beta}_1$ being the projection of $\boldsymbol{\Sigma}_X^{0.5} \boldsymbol{\beta}$ to the subspace spanned by $\boldsymbol{\Sigma}_X^{0.5} \mathbf{C}$ under the Euclidean metric) and $\boldsymbol{\beta}_2 = \boldsymbol{\beta} - \boldsymbol{\beta}_1$. Then since $\boldsymbol{\Sigma}_X^{0.5} \boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_X^{0.5} \mathbf{C}$ are both orthogonal subspaces, we have $\|\mathbf{C}^\dagger \boldsymbol{\beta}_1\| = \|(\boldsymbol{\Sigma}_X^{0.5} \mathbf{C})^\dagger (\boldsymbol{\Sigma}_X^{0.5} \boldsymbol{\beta}_1)\| \leq \|(\boldsymbol{\Sigma}_X^{0.5} \mathbf{C})^\dagger (\boldsymbol{\Sigma}_X^{0.5} \boldsymbol{\beta})\| \leq 1$, where $\dagger$ represents the pseudo inverse. Note that $\boldsymbol{\beta}_1$ and $\mathbf{C}$ have the same column space, so for any $\mathbf{z} \in \mathbb{R}^p$, we have

$$\|\boldsymbol{\beta}_1 \mathbf{z}\| \leq \|\mathbf{C}\mathbf{z}\|. \tag{7}$$

Then we proved that $\mathbf{C}$ is the solution to $\arg\min_{\mathbf{C}^T\Sigma_X\mathbf{C}=I_d} \nu^2(\mathbf{C}^T\mathbf{X},\mathbf{Y})$ asymptotically:

$$\begin{aligned}
&\nu^2(\boldsymbol{\beta}^T X, Y, \alpha)\\
&= \int |\mathrm{E}e^{i\langle t,\boldsymbol{\beta}^T X\rangle + i\langle s,Y\rangle} - \mathrm{E}e^{i\langle t,\boldsymbol{\beta}^T X\rangle}\mathrm{E}e^{i\langle s,Y\rangle}|^2 \big/ (\|t\|^{d+\alpha}\|s\|^{1+\alpha})\mathrm{d}t\mathrm{d}s\\
&= \int |\mathrm{E}e^{i\langle t,\boldsymbol{\beta}_2^T X\rangle}|^2 |\mathrm{E}e^{i\langle t,\boldsymbol{\beta}_1^T X\rangle + i\langle s,Y\rangle} - \mathrm{E}e^{i\langle t,\boldsymbol{\beta}_1^T X\rangle}\mathrm{E}e^{i\langle s,Y\rangle}|^2 \big/ (\|t\|^{d+\alpha}\|s\|^{1+\alpha})\mathrm{d}t\mathrm{d}s\\
&\le \int |\mathrm{E}e^{i\langle t,\boldsymbol{\beta}_1^T X\rangle + i\langle s,Y\rangle} - \mathrm{E}e^{i\langle t,\boldsymbol{\beta}_1^T X\rangle}\mathrm{E}e^{i\langle s,Y\rangle}|^2 \big/ (\|t\|^{d+\alpha}\|s\|^{1+\alpha})\mathrm{d}t\mathrm{d}s\\
&= \nu_n^2(\boldsymbol{\beta}_1^T X, Y, \alpha) \le \nu^2(\mathbf{C}^T X, Y, \alpha),
\end{aligned}$$

where the last step follows from Equation (7). It is easy to verify that the equality only holds when $\boldsymbol{\beta} = \mathbf{C}Q$ for some rotation matrix Q.

It remains to shows that $\nu_n^2(\mathbf{C}^T\mathbf{X},\mathbf{Y},\alpha)$ is the empirical estimate of the random variable $\nu^2(\mathbf{C}^T X, Y, \alpha)$, which means that $\nu_n^2(\mathbf{C}^T\mathbf{X},\mathbf{Y},\alpha) \xrightarrow{a.s.} \nu^2(\mathbf{C}^T\mathbf{X},\mathbf{Y},\alpha)$ (almost sure convergence) as $n \to \infty$. The result holds following the proof of Lemma 2 in the supplementary material of Zhang and Yin (2015). ■

3.1. Convergence analysis

We investigate the convergence property of the proposed algorithm in this section. In fact, the proposed algorithm generates solutions that converge to a stationary point of $F_\eta(\mathbf{C})$ as $t \to \infty$. In addition, the algorithm converges to the solution when well-initialised.

Theorem 3.2: (a) *Any accumulation point of the sequence* $\{\hat{\mathbf{C}}^{(t)}\}_{t\ge 0}$ *generated by the proposed algorithm converges is a stationary point of* $F_\eta(\mathbf{C})$ *over the Stiefel manifold.*

(b) *If in addition, the global maximiser* $\hat{\mathbf{C}}$ *it is the unique stationary point in its neighbourhood* $\mathcal{N}$*, and* $F_\eta(\mathbf{C}) - F_\eta(\hat{\mathbf{C}}) \le -c\|\mathbf{C} - \hat{\mathbf{C}}\|_F^2$ *for any* $\mathbf{C}$ *in* $\mathcal{N}$ *and some* $c > 0$*. Then when the initialisation* $\hat{\mathbf{C}}^{(0)}$ *is sufficiently close to* $\hat{\mathbf{C}}$*, the sequence* $\{\hat{\mathbf{C}}^{(t)}\}_{t\ge 0}$ *converges to* $\hat{\mathbf{C}}$*.*

Proof: (a) Due to the line search strategy in Algorithm 1, the objective value of the objective function is monotonically nondecreasing and as a result, $\nu_n^2(\hat{\mathbf{C}}^{(t)\,T}\mathbf{X},\mathbf{Y},\alpha)$ converges. Let $\tilde{\mathbf{C}}$ be any accumulation point of the sequence $\hat{\mathbf{C}}^{(t)}$, then $\nabla_{\mathbf{C}}\nu_n^2(\mathbf{C}^T\mathbf{X},\mathbf{Y},\alpha)|_{\mathbf{C}=\hat{\mathbf{C}}^{(t)}} = 0$, since otherwise the objective function will continue to increase.

(b) Since the gradient of $F_\eta(\mathbf{C})$ is continuous, $\max_{\mathbf{C}:\|\mathbf{C}-\hat{\mathbf{C}}\|_F\le\epsilon} \|F_\eta(\mathbf{C})\|$ converges to zero as $\epsilon \to 0$. As a result, we may choose $\epsilon' > 0$ such that for

$$\mathcal{N}_{\epsilon'} = \mathcal{N} \cap \{F_\eta(\mathbf{C}) - F_\eta(\hat{\mathbf{C}}) > -\epsilon'\},$$

and any $\hat{\mathbf{C}}^{(t)} \in \mathcal{N}_{\epsilon'}$, $\|\hat{\mathbf{C}}^{(t)} - \hat{\mathbf{C}}\|_F \le \sqrt{\epsilon'/c}$ and the gradient $F'_\eta(\hat{\mathbf{C}}^{(t)})$ is so small such that the next iteration $\hat{\mathbf{C}}^{(t+1)}$ remains in $\mathcal{N}$. Since the functional value $F_\eta(\hat{\mathbf{C}}^{(t)})$ is nonincreasing, $\hat{\mathbf{C}}^{(t+1)}$ lies in $\mathcal{N}_{\epsilon'}$ as well. As $\hat{\mathbf{C}}$ is the unique stationary point in $\mathcal{N}_{\epsilon'}$, part (a) implies that the algorithm converges to $\hat{\mathbf{C}}$. ■

4. Numerical studies

In this section, we perform a comparative analysis of several algorithms including the proposed robust SDR (rSDR), the SQP algorithm (Sheng and Yin 2013), the MMRN algorithm (Wu and Chen 2021) and the HSIC algorithm (Zhang and Yin 2015).

The problem in Equation (4) is nonlinear and the proposed algorithm, rSDR, needs an good initialisation. The solutions of the sliced inverse regression (SIR, Li 1991) and the directional regression (DR, Li and Wang 2007) are used in the initialisation of Algorithm 1. Let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ be two solutions of SDR obtained by SIR and DR, respectively. We select one of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ with larger dCov as our initial value of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\Sigma}}$ be the sample covariance of $\{\boldsymbol{x}\}_{i=1}^n$. The initial matrix $\mathbf{C}^{(0)} = \hat{\boldsymbol{\Sigma}}_X^{1/2}\boldsymbol{\beta}$ is evaluated in Algorithm 1.

The proposed algorithm has an parameter α which governs robustness to outliers. A smaller α usually enhances the robustness of Algorithm 1. However, an excessively small α often results in numerous local minimum values for the problem. Therefore, α is tuned through 5-fold cross-validation. The value of α is fine-tuned from $\{i/10\}_{i=1}^9$ by 5-fold CV. Specifically, we partition the datasets $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ into training and validation sets. For each α value, we apply Algorithm 1 to the training set, yielding a subspace $\boldsymbol{\beta}_\alpha$. We then assess the 0.5-dCov of the validation set. This process is repeated for all 5 folds, and the average 0.5-dCov is computed. We choose the α value associated with the highest average and execute Algorithm 1 again to derive the estimated subspace. It is important to note that if the dataset is contaminated with outliers, the validation set will also contain outliers. Traditional dCov or covariance calculations may be significantly impacted by these outliers. Therefore, opting for a more robust variance statistic is crucial. In this context, we select the 0.5-dCov as the measure for the test set.

The SQP algorithm utilises sequential quadratic programming to solve the dCov-based SDR model (equivalent to Equation (4) with $\alpha = 1$). While the SQP method performs well when the dimension (p) and sample size (n) are relatively small, it becomes computationally difficult for moderately high-dimensional settings (Wu and Chen 2021). MMRN was later proposed as an efficient alternative to solve the same model using Riemannian Newton's method. Both SQP and MMRN correspond to rSDR with $\alpha = 1$, but none of them is robust against outliers. The Hilbert–Schmidt Independence Criterion (HSIC) method (Zhang and Yin 2015) addresses the single-index SDR model ($d = 1$) by maximising the HSIC covariance between $\boldsymbol{\beta}^T\mathbf{X}$ and $\mathbf{Y}$.

In the first simulation, we compare rSDR with SQP and MMRN in both robust and non-robust settings. Our results demonstrate that rSDR with a smaller α can effectively estimate the underlying subspace and efficiently solve the SDR model. Additionally, even in the presence of outliers in the response, rSDR can still estimate the subspace accurately, while SQP and MMRN fail to do so.

In the second simulation, we explore the application of rSDR in outlier detection. By reducing the data dimension, we extend a dCor-based outlier detection method (Wang and Li 2017) to high-dimensional cases. We compare rSDR with PCA in dimensionality reduction and outlier detection to showcase the applicability of robust SDR in outlier detection.

Furthermore, we present three real data examples: the New Zealand horse mussels, cardiomyopathy microarray data and auto miles per gallon (MPG) data. In the New Zealand horse mussels dataset, we reduce the data dimension to 1 and compare rSDR with HSIC.

Table 1. The mean and standard deviation (in parentheses) of the principal angles and the running times (seconds) over 100 repetitions of SQP, MMRN and rSDR in nonrobust settings.

		SQP		MMRN		rSDR	
(n,p)	Model	Angle	Time(s)	Angle	Time(s)	Angle	Time(s)
(100,6)	A(1)	0.27(0.09)	0.16(0.12)	0.27(0.09)	0.19(0.09)	0.27(0.09)	0.17(0.13)
	A(2)	0.25(0.08)	0.13(0.13)	0.25(0.08)	0.15(0.12)	0.25(0.08)	0.18(0.13)
	B(1)	0.28(0.09)	0.10(0.02)	0.28(0.09)	0.19(0.12)	0.28(0.09)	0.20(0.14)
	B(2)	0.22(0.08)	0.11(0.04)	0.22(0.08)	0.32(0.59)	0.21(0.08)	0.23(0.18)
	C(1)	0.20(0.07)	0.24(0.32)	0.20(0.07)	0.25(0.11)	0.19(0.06)	0.08(0.05)
	C(2)	0.32(0.12)	0.14(0.24)	0.31(0.12)	0.38(0.17)	0.32(0.12)	0.08(0.05)
(500,20)	A(1)	0.24(0.04)	2.98(0.56)	0.24(0.04)	0.90(0.14)	0.24(0.04)	1.41(0.77)
	A(2)	0.23(0.04)	3.33(3.11)	0.23(0.04)	0.90(1.36)	0.23(0.04)	1.65(0.80)
	B(1)	0.24(0.04)	3.17(0.63)	0.24(0.04)	0.90(0.14)	0.24(0.04)	1.54(0.91)
	B(2)	0.19(0.03)	4.55(1.29)	0.19(0.03)	0.81(0.13)	0.18(0.03)	1.56(0.79)
	C(1)	0.16(0.03)	2.54(0.22)	0.16(0.03)	1.52(0.26)	0.17(0.03)	0.66(0.42)
	C(2)	0.25(0.04)	3.27(0.71)	0.25(0.04)	4.04(1.23)	0.28(0.05)	0.72(0.49)

Notably, HSIC is only applicable when $d = 1$, so we do not include it in other simulations or real data examples.

4.1. Simulation data

Let $\tilde{\beta}_1 = (1,0,0,\ldots,0)^T, \tilde{\beta}_2 = (0,1,0,\ldots,0)^T, \tilde{\beta}_3 = (1,0.5,1,\ldots,0)^T$ be three p-dimen sional vectors. We further rotate the vectors $\tilde{\beta}_i$ by a random rotation matrix $R_d \in SO(p)$ (the special orthogonal group of dimension p), i.e. $\beta_i = R_d^\top \tilde{\beta}_i$. We consider the following three models

(A) $Y = (\beta_1^T X)^2 + (\beta_2^T X) + 0.1\epsilon$,
(B) $Y = \text{sign}(2\beta_1^T X + \epsilon_1) \times \log|2\beta_2^T X + 4 + \epsilon_2|$,
(C) $Y = \exp(\beta_3^T X)\epsilon$,

where $X \in \mathbb{R}^p$ follows from (1) $\mathcal{N}(0, \mathbf{I})$ and (2) $U[-2,2]^p$ and $\epsilon, \epsilon_1, \epsilon_2$ are standard normal distributed. We analyse the principal angles between the true subspace $\boldsymbol{\beta}$ and the estimated subspace $\hat{\boldsymbol{\beta}}$ obtained using different SDR methods, namely rSDR, MMRN and SQP. To further investigate the robustness of these methods, we introduce additional noise by adding the response with a value of $50 \times \mathbf{1}^T X$ with a probability of 0.1. We then calculate the principal angles between the true subspace and the estimated subspaces in this robust setting. Both simulation scenarios are conducted for two settings: $(n,p) = (100,6)$ and $(n,p) = (500,20)$. We repeat the simulations 100 times and report the mean and standard deviation of the principal angles for both the non-robust and robust cases in Tables 1 and 2, respectively. It is worth noting that the underlying subspace for model (A) and (B) is represented by $\boldsymbol{\beta} = [\beta_1, \beta_2]$, resulting in a value of $d = 2$. On the other hand, the underlying subspace for model (C) is represented by $\boldsymbol{\beta} = \beta_3$, resulting in a value of $d = 1$.

From Table 1 we observe that rSDR performs better than MMRN and SQP in model (A) and (B) even without outliers. MMRN converges faster than SDR in model (A) and (B). When $(n,p) = (500,20)$, rSDR and MMRN are faster than SQP. Table 2 reports the principal angles and execution time of the three estimators in the scenario where the outliers present. Table 2 shows that the principal angles between the true subspace and the

Table 2. The mean and standard deviation (in parentheses) of the principal angle, and the running time (seconds) over 100 repetitions of SQP, MMRN and rSDR in robust settings.

		SQP		MMRN		rSDR	
(n,p)	Model	Angle	Time(s)	Angle	Time(s)	Angle	Time(s)
(100,6)	A(1)	0.51(0.28)	0.20(0.21)	0.49(0.27)	0.28(0.20)	0.32(0.12)	0.19(0.13)
	A(2)	0.45(0.28)	0.13(0.12)	0.44(0.27)	0.26(0.19)	0.27(0.11)	0.17(0.12)
	B(1)	0.52(0.26)	0.13(0.08)	0.51(0.26)	0.32(0.26)	0.33(0.12)	0.18(0.13)
	B(2)	0.42(0.22)	0.12(0.07)	0.42(0.22)	0.25(0.22)	0.24(0.08)	0.21(0.17)
	C(1)	0.39(0.24)	0.19(0.21)	0.38(0.23)	0.32(0.16)	0.26(0.10)	0.08(0.06)
	C(2)	0.47(0.23)	0.17(0.27)	0.46(0.22)	0.51(0.25)	0.40(0.16)	0.09(0.06)
(500,20)	A(1)	0.82(0.30)	4.12(1.29)	0.82(0.30)	2.92(1.47)	0.25(0.04)	1.47(0.87)
	A(2)	0.93(0.42)	16.26(42.21)	0.92(0.42)	6.85(13.70)	0.24(0.04)	1.43(0.85)
	B(1)	0.91(0.35)	4.20(1.48)	0.90(0.35)	3.36(2.23)	0.26(0.04)	1.74(0.92)
	B(2)	0.60(0.36)	5.32(2.26)	0.60(0.36)	2.84(3.54)	0.19(0.03)	1.64(0.99)
	C(1)	0.35(0.14)	3.02(0.34)	0.35(0.14)	4.41(2.15)	0.22(0.04)	0.63(0.46)
	C(2)	0.89(0.28)	4.53(1.01)	0.85(0.29)	13.39(6.27)	0.35(0.07)	0.88(0.61)

estimated subspace produced by rSDR are smaller than MMRN and SQP which implies that rSDR is more robust. Moreover, rSDR converges faster than MMRN and SQP in most settings; particularly in model (C).

4.2. Outlier detection simulation studies

Our proposed SDR method can be effectively utilised for outlier detection. Wang and Li (2017) introduced a novel outlier detection measure based on the distance correlation (dCor) given by

$$\mathcal{D}_i(\mathbf{X},\mathbf{Y}) = \frac{1}{p}\sum_{k=1}^{p}\left(\mathrm{dCor}(\mathbf{X}_k,\mathbf{Y}) - \mathrm{dCor}(\mathbf{X}_k^{(i)},\mathbf{Y}^{(i)})\right)^2, \tag{8}$$

where $\mathrm{dCor}(\mathbf{X}_k,\mathbf{Y})$ represents the dCor between the kth predictor and the response $\mathbf{Y}$. The dCor between $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathrm{dCor}^2(\mathbf{X},\mathbf{Y}) = \frac{\mathrm{dCov}^2(\mathbf{X},\mathbf{Y})}{\sqrt{\mathrm{dCov}^2(\mathbf{X},\mathbf{X})\mathrm{dCov}^2(\mathbf{Y},\mathbf{Y})}}.$$

It is evident that if the ith data point $(\mathbf{X}_k^{(i)},\mathbf{Y}^{(i)})$ exhibits a high value of the measure $\hat{\mathcal{D}}_i$, it is more likely to be an outlier observation. The method employs a bootstrap procedure to determine the threshold $\hat{F}_\gamma$. At a given significance level γ, the ith observation is identified as an outlier if $\hat{\mathcal{D}}_i > \hat{F}_\gamma$, where $\hat{F}_\gamma$ represents the upper γth quantile of the cumulative distribution function of $\mathcal{D}_i$ under the null hypothesis. Specifically, a bootstrap sample $\mathcal{D}_i^{[b]}$ is formed by drawing with replacement from $1,\ldots,n$, denoted as $i_{(1)}^{[b]},\ldots,i_{(n)}^{[b]}$, and an estimator $\hat{\mathcal{D}}_i^{[b]}$ is computed for each sample. The threshold $\hat{F}_\gamma$ is determined by calculating the upper γth quantile of the cumulative distribution function of $\hat{\mathcal{D}}_i^{[b]}$.

The algorithm proposed by Wang and Li (2017), which is based on the outlier detection measure defined in Equation (8), involves calculating the covariance distance between $\mathbf{X}$ and $\mathbf{Y}$ in each dimension and with the removal of each sample. As a result, its computational complexity is $O(pn^3)$, where the computation of dCov requires calculating pairwise

distances between the columns of $\mathbf{X}$ and $\mathbf{Y}$. A natural approach to enhance their method is to reduce the dimensionality of the dataset $\mathbf{X}$. Their method can be naturally extended to detect outlier locations by computing

$$\mathcal{D}_i(\bar{\mathbf{X}}, \mathbf{Y}) = \frac{1}{p}\sum_{k=1}^{p}\left(\mathrm{dCor}(\bar{\mathbf{X}}_k, \mathbf{Y}) - \mathrm{dCor}(\bar{\mathbf{X}}_k^{(i)}, \mathbf{Y}^{(i)})\right)^2, \tag{9}$$

where $\bar{\mathbf{X}} \in \mathbb{R}^{d\times n}$ is the d-dimensional data obtained by dimension reduction. Nevertheless, the conventional approach to dimension reduction is unsuitable in the presence of outliers. Therefore, we employ the robust SDR as a means to both reducing the data's dimensionality and identifying outlier positions. For the sake of comparison, we also implement principal component analysis (PCA) (Wold et al. 1987) for dimension reduction.

We consider an autoregressive correlation structure with $\mathbf{\Sigma} = (\rho_{j,k})_{p\times p} = 0.5^{|j-k|}$ and generate the data as follows: X_i follows a multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma})$, and the linear model is defined as $Y_i = X_i\beta + \epsilon_i$, where $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)^T$ and $\epsilon_i \sim \mathcal{N}(0, 1)$. We have a total of $n = 100$ samples, and among them there are 10 outliers. The outliers are generated using $\kappa_i = X_i\gamma$, where $\gamma = (0, 0, 0, 0, 0, 1, 1, \ldots)$. We did four sets of simulations for various values of $p = 200$, 400, 800, 1000. To test the hypothesis of whether the ith observation is influential or not, we employ a bootstrap procedure and utilise a threshold rule to determine whether an individual is an outlier. We evaluate the performance of this outlier identification procedure by comparing the receiver operating characteristic (ROC) curves.

The ROC curves are depicted in Figure 1. In the figure, the curve labelled as 'PCA-2' represents the ROC curve generated by $\bar{\mathbf{X}}^{PCA}$ with a dimensionality of $d = 2$, while the curve labelled as 'rSDR-0.2-2' corresponds to the curve produced by $\bar{\mathbf{X}}^{DR}$ with $\alpha = 0.2$ and $d = 2$. Similarly, the remaining labels follow similar settings. It can be observed that the curves generated by rSDR with $d = 3$ consistently surpass those produced by rSDR with $d = 2$, and both outperform the curves generated by PCA. This suggests that the proposed rSDR method effectively captures the underlying structure of the data, and the resulting transformed data $\bar{\mathbf{X}}^{DR}$ can be utilised for outlier detection. Notably, despite the true subspace being two-dimensional, $\bar{\mathbf{X}}^{DR}$ with $d = 3$ outperforms its two-dimensional counterpart. We speculate that the higher dimensionality preserves more information due to the presence of outliers.

4.3. Real data example: New Zealand horse mussels

A sample of 201 horse mussels (Modiolus modiolus) was collected at 5 sites in the Marlborough Sounds at the Northeast of New Zealand's South Island and this dataset was discussed by Cook (2009). The response variable is muscle mass M, the edible portion of the mussel, in grams. The quantitative predictors are all related to the characteristics of the mussel shells: shell width W (in mm), shell height H (in mm), shell length H (in mm) and shell mass S (in grams).

To process the data, a nonlinear transformation of the predictors was recommended by Cook (2009) as $X = (L, W^{0.36}, S^{0.11})$. Each column of the data X is further standardised by $\tilde{X} = (\frac{L-\hat{\mu}_L}{\hat{\sigma}(L)}, \frac{W^{0.36}-\hat{\mu}_{W^{0.36}}}{\hat{\sigma}(W^{0.36})}, \frac{S^{0.11}-\hat{\mu}_{S^{0.11}}}{\hat{\sigma}(S^{0.11})})$ where $\hat{\mu}_\cdot$ is the sample mean and $\hat{\sigma}(\cdot)$ is the sample

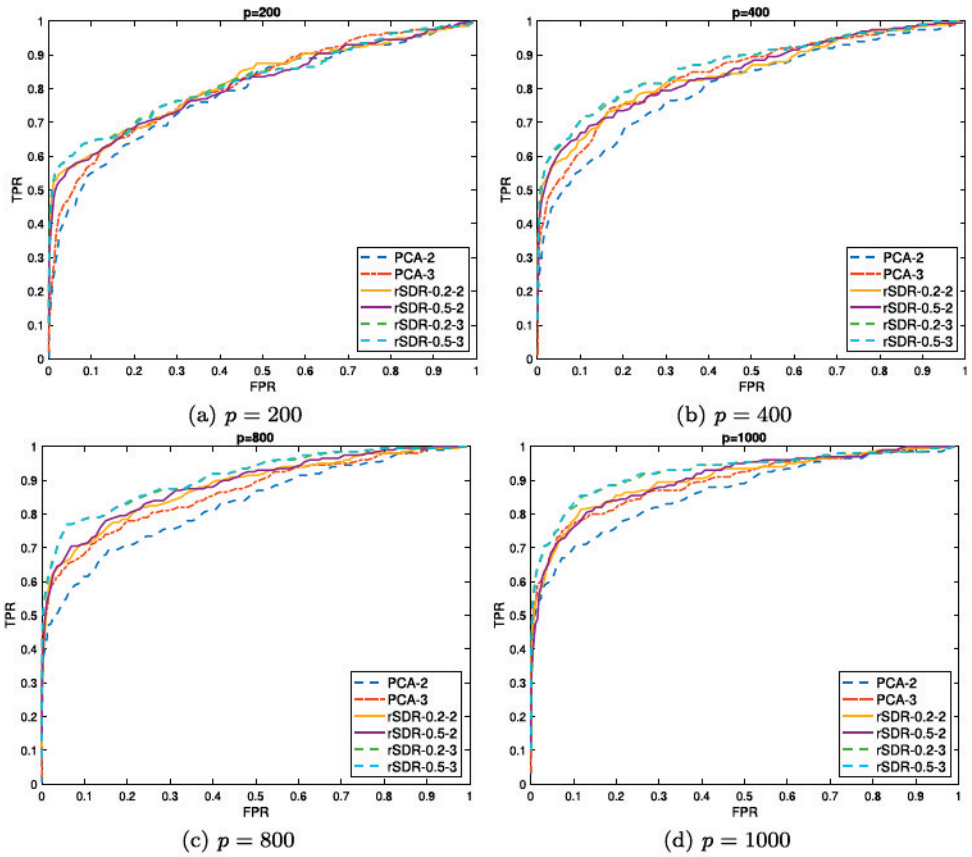
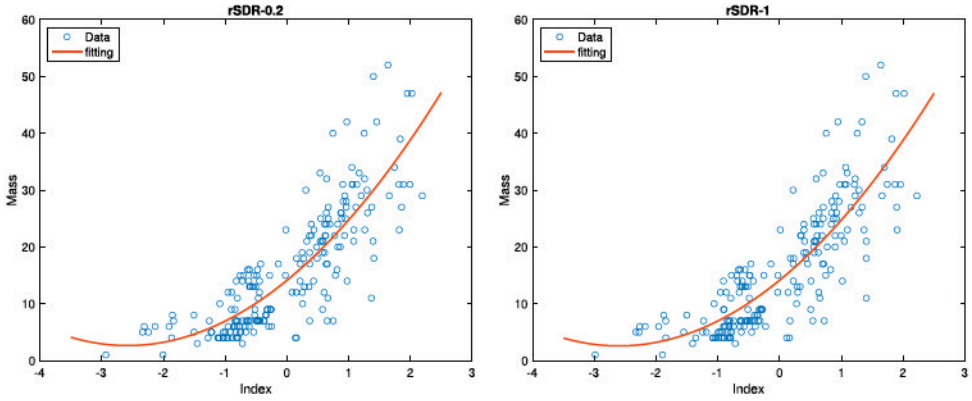


Figure 1. ROC curves of outlier detection. The proposed robust SDR method with $\alpha = 0.5$ with projection dimension 3 has the highest ROC in these four simulation sample size settings. (a) $p = 200$, (b) $p = 400$, (c) $p = 800$ and (d) $p = 1000$.

standard deviation, since L is on a larger scale than the other predictors. Consequently, the predictors will have mean 0 and variance 1. The rSDR model with $d = 1$ would be appropriate to model this dataset, as shown in Figure 2, where we fit two second-degree polynomial regression models of the single index $\hat{\beta}^T\tilde{X}$ by rSDR with $\alpha = 0.2$ and $\alpha = 1$. We compare our method rSDP with $\alpha = 0.2$ and $\alpha = 1$, SQP and the Hilbert–Schmidt Independence Criterion (HSIC) method, proposed by Zhang and Yin (2015) for solving the special case of the SDR model, namely $d = 1$. Table 3 provides the estimated bases β from these four methods. The estimates of SDR with $\alpha = 1$ and SQP are similar, and this result is expected since SQP and rSDR with $\alpha = 1$ solve the same model with different algorithms. The estimated $\hat{\beta}$ from all four methods indicate that the standardised shell mass predictor, $\tilde{X}_3$, is more significant than the other two predictors while the rSDR with $\alpha = 0.2$ produces a smaller value in the coefficient of $\tilde{X}_3$. However, rSDR with $\alpha = 0.2$ produces a model with a slightly larger R-squared value than the other methods, which implies a better fit of the dataset.



(a) rSDR-0.2 (b) rSDR-1

Figure 2. The second-degree polynomial fitting of the single-index model in the New Zealand Horse Mussels data using rSDR ($\alpha = 0.2$) (left) and rSDR ($\alpha = 1$) (right). (a) rSDR-0.2 and (b) rSDR-1.

Table 3. Estimated bases $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]^T \in \mathbb{R}^3$ of the central subspace in the New Zealand Horse Mussels data from various methods and their adjusted R-squared values.

Method	rSDR ($\alpha = 0.2$)	rSDR ($\alpha = 1$)	HSIC	SQP
$\hat{\beta}_1$	0.2871	0.1832	0.1897	0.1831
$\hat{\beta}_2$	0.0872	−0.0270	−0.0604	−0.0269
$\hat{\beta}_3$	0.6391	0.8510	0.9800	0.8509
Adjusted R-squared	0.7026	0.6979	0.6962	0.6979

4.4. *Real data example: cardiomyopathy microarray data*

The cardiomyopathy microarray dataset consists of 30 samples and 6319 predictors, originally used by Segal et al. (2003) to evaluate regression-based approaches for microarray analysis. The focus of many researchers, Zou and Yuan (2008) and Li et al. (2012), has been to investigate the relationship between the overexpression of a G protein-coupled receptor (Ro1) in mice and the 6319 associated genes. However, due to the high dimensionality of the data compared to the limited number of samples, the sample covariance matrix is not invertible. To address this issue, several methods have been proposed, including SIS (Sure Independence Screening, Fan and Lv (2008)), DCSIS (Distance Correlation SIS, Li et al. (2012)), BCSIS (Ball Correlation SIS, Pan et al. (2019)) and SDRLS (Sequential Dimension Reduction for Large p Small n problem, Yin and Hilafu (2015)). While SIS, DCSIS and BCSIS are feature screening methods that rank predictors based on a utility measure, they may not be robust against outliers. Specifically, a set of predictors $\mathcal{A} = \{i \mid U(X_i, Y) > \tau, i = 1, \ldots, n\}$ is determined for some threshold τ and pre-selected utility measure U. SDRLS takes a different approach. SDRLS partitions the data set into $X = [X_1, X_2]$ with $\dim(X_1) < n$ and applies the SDR model on $(X_1, [X_2, Y])$ to obtain $R(X_1)$. The dimension of $R(X_1)$ is chosen some integer that is smaller than $\dim(X_1) < n$ and thus a new predictor $[R(X_1), R_2]$ is obtained with a smaller dimension. SDRLS iteratively repeats this process to achieve a dimension smaller than the number of samples.

Table 4. Adjusted R-squared and F-value of models from SQP, rSDR in Cardiomyopathy Microarray dataset.

Adjusted R-squared	rSDR ($\alpha = 0.2$)	rSDR ($\alpha = 0.5$)	rSDR ($\alpha = 1$)
Linear	0.826	0.817	0.804
Nonlinear	0.882	0.871	0.867
F-value	rSDR ($\alpha = 0.2$)	rSDR ($\alpha = 0.5$)	rSDR ($\alpha = 1$)
Linear	70.1	65.8	60.6
Nonlinear	44.4	40.3	38.9

In this experiment, we utilised the SDRLS method to reduce the dimensionality of the cardiomyopathy microarray data and assess the rSDR against heavy-tailed predictors. The final dimension of the dataset was reduced to $p = 19$, while the dimension of the central subspace was set to $d = 2$. The central subspace is denoted as $\boldsymbol{\beta} = [\beta_1, \beta_2]$. Indexes derived from this reduction were obtained by projecting the processed cardiomyopathy microarray dataset $\mathbf{X}$ onto the subspaces: $\mathbf{Z}_1 = \beta_1^T\mathbf{X}$ and $\mathbf{Z}_2 = \beta_2^T\mathbf{X}$. We performed linear and nonlinear regression to model the response variable 'Ro1' using predictors $\mathbf{Z}_1$ and $\mathbf{Z}_2$. In the nonlinear model, we introduced squared terms $(\mathbf{Z}_1^2, \mathbf{Z}_2^2)$ and an interaction term $(\mathbf{Z}_1 \times \mathbf{Z}_2)$ in addition to the linear model. The regression results are presented in Table 4. The findings demonstrate that our proposed method, rSDR, with a smaller value of α, outperforms the non-robust version ($\alpha = 1$) in both linear and nonlinear models.

4.5. Real data example: auto MPG data

We also employ the fuel economy MPG data to illustrate the advantage of our rSDR method. The dataset contains city-cycle fuel consumption in MPG and seven predictors: cylinders, displacement, horsepower, weight, acceleration, model year and origin. As suggested in Sheng and Yin (2016), we avoid using 'origin', because it correlates with 'cylinders' closely. Missing values are deleted, and 392 observations are left for study. In order to investigate the city-cycle fuel consumption in miles per gallon, we assume that this dataset fits a sufficient dimension reduction model. As shown in Figure 3, there exist outliers in 'horsepower' and 'acceleration'. 'cylinders' and 'displacement' that are not normally distributed. Therefore, rSDR is appropriate for this data set.

Following the suggestion of Sheng and Yin (2016), we use the dimension $d = 2$ of the central subspace. Let the subspace be $\boldsymbol{\beta} = [\beta_1, \beta_2]$ and the auto MPG data be denoted as $\mathbf{X}$ where each column of $\mathbf{X}$ is centred and scaled to make the variance as 1. The following procedures are similar to those done in the cardiomyopathy microarray data. The indexes are derived by rSDR, after the linear and nonlinear regression models are constructed to measure the goodness of fit of the two pair of indexes to 'mpg'. In Table 5, the adjusted R-squared and F-value of the linear model produced by rSDR with $\alpha = 0.2$ are larger than other non-robust models however it does not show much superior in the nonlinear regression model (Figure 4).

5. Discussion

In this article, the proposed rSDR using α-dCov is robust against outliers in both the response and predictors. Further, the proposed manifold-learning estimation method is

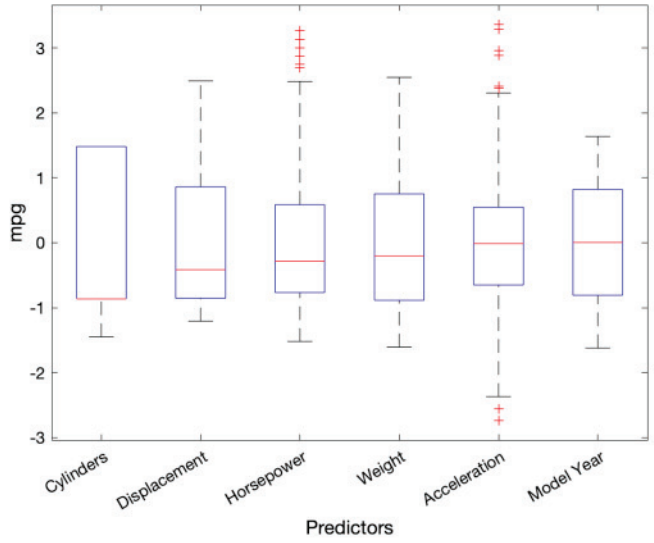


Figure 3. Boxplots of predictors in the auto mpg data set. Some predictors such as 'horsepower' and 'acceleration' have outlying observations.

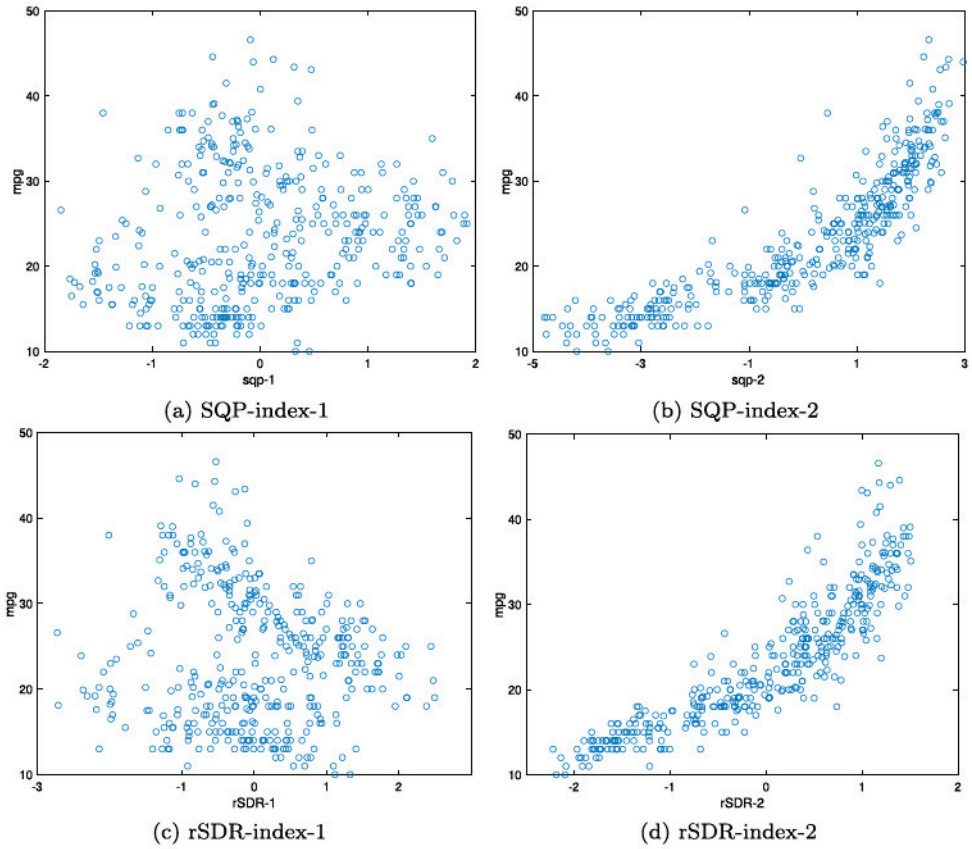


Figure 4. Scatter plots of 'mpg' versus the indexes produced by SQP and rSDR ($\alpha = 0.2$). (a) SQP-index-1, (b) SQP-index-2, (c) rSDR-index-1 and (d) rSDR-index-2.

Table 5. Adjusted R-squared and F-value of models from SQP, rSDR in MPG dataset.

Adjusted R-squared	rSDR ($\alpha = 0.2$)	rSDR ($\alpha = 0.5$)	rSDR ($\alpha = 1$)
Linear	0.807	0.806	0.804
Nonlinear	0.850	0.853	0.845
F-value	rSDR ($\alpha = 0.2$)	rSDR ($\alpha = 0.5$)	rSDR ($\alpha = 1$)
Linear	817	816	807
Nonlinear	444	456	427

less sensitive to the choice of the initial estimators. Both simulation and real-world data applications show that the proposed method outperforms the existing methods. The proposed method does not suffer from multicollinearity which could impact the performance of the traditional SDR methods in high-dimensional data analysis. Simulation and real-world data studies show its advantages in terms of computational efficiency and robustness against outliers.

Acknowledgments

The authors would like to thank the Editor, the Associate Editor and the reviewers for their constructive and insightful comments that greatly improved the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by NSF grants (DMS-1924792, DMS-2318925 and CNS-1818500).

References

Absil, P., Mahony, R., and Sepulchre, R. (2009), *Optimization Algorithms on Matrix Manifolds*, Princeton, NJ: Princeton University Press. https://books.google.com/books?id=NSQGQeLN3NcC

Absil, P.-A., and Malick, J. (2012), 'Projection-like Retractions on Matrix Manifolds', *SIAM Journal on Optimization*, 22(1), 135–158.

Cook, R.D. (2009), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, Hoboken, NJ: John Wiley & Sons.

Cook, R.D., and Weisberg, S. (1991), 'Discussion of Sliced Inverse Regression for Dimension Reduction', *Journal of the American Statistical Association*, 86(414), 328–332.

Dalmau-Cedeno, O., and Oviedo, H (2017), 'A Projection Method for Optimization Problems on the Stiefel Manifold', in *Mexican Conference on Pattern Recognition*, pp. 84–93.

Fan, J., and Lv, J. (2008), 'Sure Independence Screening for Ultrahigh Dimensional Feature Space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849–911.

Hall, P., and Li, K.-C. (1993), 'On Almost Linearity of Low Dimensional Projections From High Dimensional Data', *The Annals of Statistics*, 21(2), 867–889.

Li, K.-C. (1991), 'Sliced Inverse Regression for Dimension Reduction', *Journal of the American Statistical Association*, 86(414), 316–327. doi:10.1080/01621459.1991.10475035

Li, B., and Wang, S. (2007), 'On Directional Regression for Dimension Reduction', *Journal of the American Statistical Association*, 102(479), 997–1008.

Li, R., Zhong, W., and Zhu, L. (2012), 'Feature Screening via Distance Correlation Learning', *Journal of the American Statistical Association*, 107(499), 1129–1139.
Pan, W., Wang, X., Xiao, W., and Zhu, H. (2019), 'A Generic Sure Independence Screening Procedure', *Journal of the American Statistical Association*, 114(526), 928–937.
Segal, M.R., Dahlquist, K.D., and Conklin, B.R. (2003), 'Regression Approaches for Microarray Data Analysis', *Journal of Computational Biology*, 10(6), 961–980.
Sheng, W., and Yin, X. (2013), 'Direction Estimation in Single-index Models via Distance Covariance', *Journal of Multivariate Analysis*, 122, 148–161.
Sheng, W., and Yin, X. (2016), 'Sufficient Dimension Reduction via Distance Covariance', *Journal of Computational and Graphical Statistics*, 25(1), 91–104.
Székely, G.J., and Rizzo, M.L. (2009), 'Brownian Distance Covariance', *The Annals of Applied Statistics*, 3(4), 1236–1265.
Székely, G.J., and Rizzo, M.L. (2014), 'Partial Distance Correlation with Methods for Dissimilarities', *The Annals of Statistics*, 42(6), 2382–2412. doi:10.1214/14-AOS1255
Székely, G.J., Rizzo, M.L., and Bakirov, N.K. (2007), 'Measuring and Testing Dependence by Correlation of Distances', *The Annals of Statistics*, 35(6), 2769–2794. doi:10.1214/009053607000000505
Wang, T., and Li, Z. (2017), 'Outlier Detection in High-dimensional Regression Model', *Communications in Statistics-Theory and Methods*, 46(14), 6947–6958.
Wen, Z., and Yin, W. (2013), 'A Feasible Method for Optimization with Orthogonality Constraints', *Mathematical Programming*, 142(1-2), 397–434. doi:10.1007/s10107-012-0584-1
Wold, S., Esbensen, K., and Geladi, P. (1987), 'Principal Component Analysis', *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37–52.
Wu, R., and Chen, X. (2021), 'Mm Algorithms for Distance Covariance Based Sufficient Dimension Reduction and Sufficient Variable Selection', *Computational Statistics & Data Analysis*, 155, 107089.
Xia, Y. (2008), 'A Multiple-index Model and Dimension Reduction', *Journal of the American Statistical Association*, 103(484), 1631–1640.
Xia, Y., Tong, H., Li, W.K., and Zhu, L.-X. (2002), 'An Adaptive Estimation of Dimension Reduction Space', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3), 363–410.
Yin, X., and Cook, R.D. (2005), 'Direction Estimation in Single-index Regressions', *Biometrika*, 92(2), 371–384.
Yin, X., and Hilafu, H. (2015), 'Sequential Sufficient Dimension Reduction for Large P, Small N Problems', *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 77(4), 879–892.
Yin, X., and Li, B. (2011), 'Sufficient Dimension Reduction Based on An Ensemble of Minimum Average Variance Estimators', *The Annals of Statistics*, 39(6), 3392–3416. doi:10.1214/11-AOS950
Yin, X., Li, B., and Cook, R.D. (2008), 'Successive Direction Extraction for Estimating the Central Subspace in a Multiple-index Regression', *Journal of Multivariate Analysis*, 99(8), 1733–1757.
Zhang, J., and Chen, X. (2019), 'Robust Sufficient Dimension Reduction via Ball Covariance', *Computational Statistics & Data Analysis*, 140, 144–154.
Zhang, N., and Yin, X. (2015), 'Direction Estimation in Single-index Regressions via Hilbert-Schmidt Independence Criterion', *Statistica Sinica*, 25(2), 743–758.
Zhu, R., Zhang, J., Zhao, R., Xu, P., Zhou, W., and Zhang, X (2019), 'orthoDr: Semiparametric Dimension Reduction via Orthogonality Constrained Optimization', *The R Journal*, 11/2.
Zou, H., and Yuan, M. (2008), 'Regularized Simultaneous Model Selection in Multiple Quantiles Regression', *Computational Statistics & Data Analysis*, 52(12), 5296–5304.